4.4.7 Assessment of Model Predictability. Because it is unlikely that there will be sufficient structure-activity data to uniquely define a model at atomic resolution in competition with crystallography, justification for model building must come from its potential predictive power and possible insight into the receptor-drug interaction before detailed three-dimensional information from either crystal structure or NMR studies. Certainly, the questions regarding the ability of a proposed drug to bind to the active site without steric conflict with the receptor can be addressed by the methods outlined above in a qualitative manner. The resolution of our receptor models is too crude, however, to subject them to molecular mechanics estimates of affinities. There are alternative paradigms, however, based on pattern recognition techniques in which a set of analogs and their activities are used, along with their physicochemical parameters, to generate a mathematical model that relates the values of the physicochemical parameters for a given analog with its activity. One such paradigm is comparative molecular field analysis (CoMFA), which combines the three-dimensional electrostatic and steric fields surrounding the analogs with powerful statistical techniques, partial least squares (PLS) (477) and cross-validation, to generate predictive models if a set of orientation rules are available for aligning the molecules for comparison and prediction. Alternative methods for assessing similarity and their use in QSAR schemes have been compared (215) with CoMFA. Another approach is the use of neural nets that learn to "see" patterns in much the same way as our own nervous system processes information. Examples of the use of this pattern-recognition approach include classification of mechanism of action for cancer chemotherapy (478) and QSAR studies of DHFR inhibitors (479, 480) and carboquinones (481). Machine learning has also been applied (482) to the QSAR problem. Trimethoprim analogs were successfully analyzed for their inhibition of DHFR and similar results to the original Hansch results were obtained. It is not clear that this paradigm could be applied to noncongeneric series, at least as outlined.

What appears crucial to such studies is the choice of training set, which encompasses as much of parameter space as one is likely to use in the predictive mode as well as tests of the predictive ability of resulting models. Given that one is dealing with a situation in which the number of variables is larger (often several times) than the number of observations, linear regression models are not applicable because chance correlations are highly probable. The use of cross-validation allows selection of correlations that are predictive in a self-consistent manner within the training set. This does not mean to imply that such internally self-consistent models have predictive power outside of the training set, or extremely close congeners.

DePriest et al. (483, 484) applied the CoMFA methodology to a series of 68 ACE (angiotensin-converting enzyme) inhibitors representing 28 different chemical classes. Through use of the binding-site geometry determined by Mayer et al. (397), a CoMFA model with a statistically significant cross-validated R^2 and considerable predictive ability for inhibitors outside of the training set was derived. Because the geometry of the ACE inhibitors was determined computationally by an active-site analysis rather than experimentally, a comparison of the results of the ACE series against thermolysin inhibitors, for which there were crystallographic data to explicitly define the binding-site geometry and the resulting alignment rules, was made, given that thermolysin is also a zinc-containing metallopeptidase with numerous similarities between ACE and thermolysin. Their results give strong support to both the Active Analog Approach (398) used to define the alignment rule for the ACE series and the CoMFA methodology itself. In the absence of an experimentally known active-site geometry, correlations were derived that explain as much as 84% of the variance in activities among a set of 68 diverse ACE inhibitors by use of CoMFA steric and electrostatic potentials plus a zinc indicator variable (Fig. 3.50). If the set of 68 ACE inhibitors was divided into three classes and correlations are derived for each class, CoMFA parameters alone explain 79–99% of the variance in activities. It was notable that statistically significant correla-